MULTI-RESOLUTION FRONT-END FOR NOISE ROBUST SPEECH RECOGNITION
Ramalingam Hariharan, Imre Kiss, Olli Viikki, Jilei Tian

Speech and Audio Systems Laboratory, Nokia Research Center, Tampere, Finland
Email: {ramalingam.hariharan, imre.kiss, olli.viikki, jilei.tian, juha.hakkinen}@nokia.com

ABSTRACT
This paper proposes a new feature extraction approach for noise robust speech recognition. The recent work in multi-band and missing feature theory based Automatic Speech Recognition (ASR) has shown that sub-band processing of speech has certain advantages over the conventional full-band technique. In multi-band ASR, different frequency sub-bands are usually decoded independently and a final recognition result is obtained by combining different frequency channels at some temporal level. Since it is not straightforward to determine the optimal combination level, we propose that different sub-band parameters need to be collected into a single feature vector for decoding. As the full-band parameters still carry important information for classification, we suggest that full-band features need to be included in the final feature vector. Our third observation is that the use of PCA transform for de-correlating log-spectral features provides better recognition performance than DCT. The experimental results show that the proposed front-end provides 36.2% improvement in performance over the conventional full-band technique.

1. INTRODUCTION
Noise robust feature extraction is probably one of the greatest challenges in ASR. Starting from the early days of speech recognition, researchers have well understood that noise immune speech parameterisation is a must when developing high performance speech recognition systems. Despite the large amount of efforts put on developing robust feature extraction methods, the state-of-the-art feature extraction algorithms are unfortunately not capable of providing fully invariant features. The representation of speech tends to change according to the characteristics of an acoustic environment. This fact makes the current ASR systems to fail if there is a large mismatch between the training and testing conditions.

Mel Frequency Cepstral Coefficients (MFCCs) have been used conventionally as the front-end for many of the speech recognition systems. Even though the MFCC representation of speech has been standardised for cellular systems [10], it does not cope very well with noisy speech. Many alternate feature extraction methods have been proposed over the years that provide better performance in noisy conditions. A set of perceptual linear prediction (PLP) coefficients, which tries to incorporate some of the features of the human auditory mechanism has been proposed. Together with RASTA filtering, these features have been shown to perform better in noisy conditions [2]. In [3], an auditory model based front-end that relied on short-term adaptation proved to be more noise robust than conventional MFCC features for name dialling application.

Sub-band based speech recognition has attracted a lot of interest in recent years. One of the prime motivations is attributed to [4], where it is argued that the human auditory system processes the features from different sub-bands independently and the merging is done at some point of processing to produce the final decision. It is also understandable that noise usually corrupts only parts of the full frequency band. Since feature analysis is done independently for each part of the spectrum, it does not spread noise over all features, but only to those computed from the corrupted part of the spectrum. Hence the multi-band features can be claimed to be more robust to ambient noise.

Most of the work in multi-band ASR has focused on recombining the likelihoods produced by the different sub-band recognisers. The work done by Hermansky, Bourlard et al. [8,9] focused on combining the likelihood scores at some recombination level such as phone, HMM state or word etc. Recent work has also focused on recombining the multiple inputs at the feature level [6]. Even though the systems proved to be noise robust in narrow band noise conditions, most of the papers did not show any substantial increase in overall recognition accuracies. In [5], it was noticed that while multi-band speech recognition performed better than full-band recognition for certain kinds of noise, it was vice-versa for some other noise types.

We here propose a method where we combine both the full-band and various sub-band features into a single feature vector. This approach is able to combine the advantages produced by both the full-band and different sub-bands in different noise conditions. Experimental results prove the viability of our approach. The remainder of this paper is organised as follows. The next section briefly outlines the multi-resolution front-end approach. It is followed by experimental evaluation and the results obtained. The paper concludes with a summary of the work.

2. MULTI-RESOLUTION FRONT-END
In spite of some promising results achieved in sub-band based ASR, we believe in the importance of performing full-band processing when computing feature vectors for recognition. The major drawback of pure sub-band based approach is that the information on the correlation between various sub-bands is lost. This may be one explanation why sub-band systems have not provided consistent performance improvements over similar full-band ASR systems. Therefore, we first suggest that the full-band coefficients should not be ignored, but they should be combined with sub-band features for maximising the recognition accuracy.

It is also not trivial to decide at which temporal level the combination of sub-band features should be carried out. To avoid the use of some cumbersome combination strategy, we also suggest that sub-band combination should always be done at the feature vector level before decoding. As a main advantage of sub-band ASR, we see the fact that independent processing of
sub-band features does not spread the noise across all cepstral coefficients after linear de-correlation transformation, such as Discrete Cosine Transformation (DCT). By performing the feature analysis independently for each sub-band and combining cepstral coefficients of different sub-bands, the final feature vector is constructed, which has always some uncorrupted features available for recognition.

2.1 Multi-Resolution in the Spectral Domain

Figure 1 shows the different blocks of the proposed feature extraction algorithm. In addition to the full-band processing of the FFT spectrum, the spectrum is also processed in separate parallel blocks. The FFT output spectrum of each frame in each of these blocks is subdivided into sub-band streams, with $M_i$ ($M_i \geq 1$) FFT bins in each sub-band as shown in Figure 2. The frequency (sub-band) division into streams in each block is different. In this way, different set of sub-bands, with different frequency sub-divisions, could be independently produced in each block and added to the full-band feature vectors. Further processing in each of these streams to produce features vectors is shown in Figure 3.

As shown in Figure 3, a Principal Component Analysis (PCA) transform computed from a large number of utterances of the TIMIT database, is used as the linear transform, instead of the standard DCT, to produce the cepstral coefficients. By using PCA, better de-correlation of the subband features is achieved. It can also be observed from Figure 4, that signal co-variances in different sub-bands (also compared to full-band, which is not shown in the figure) are different, hence the use of separate transforms is justified. Experimental results in Section 3 show that the "cepstral" coefficients generated by PCA produced better recognition performance compared to the use of DCT cepstral coefficients.

The final feature vector comprises the features from the different blocks, which contain independent sub-band features, and the stream that produces the conventional full-band features, together with their time derivatives. These features are further normalised to have similar parameter statistics in all noise conditions. A simple recursive cepstral domain normalisation is carried out on the time trajectories of individual features, so that their short-term means and variances are normalised to zero and one respectively, irrespective of the environment [11].

3. EXPERIMENTAL EVALUATION

3.1 Database and Settings

The database provided by the AURORA ETSI STQ WI 008 front-end standardisation committee has been used for the work here. The database consists of TIDIGITS database pre-mixed with four different kinds of noise – exhibition hall (called N1), babble noise (N2), suburban train (N3) and moving car noise (N4) at Signal-to-Noise Ratios (SNRs) of 20, 15, 10, 5, 0 and –5
dB (in addition to the clean speech). The results with data using clean speech and at SNR of -5 dB have not been used to compute the averages here as per the AURORA draft specification. The training set consists of 422 different utterances mixed with the above noise SNRs of 20 dB, 15 dB, 10 dB and 5 dB besides the clean utterances. There is a separate set of 4000 test utterances (1000 separate utterances for each type) for recognition [1].

Training and recognition have been carried out using the HTK tool kit [7] as per the guidelines issued in [1]. Eighteen state, 3-mixture whole word models are trained from the training data. Separate 5-state, 6-mixture silence model and 3-state, 6-mixture inter-word silence models are also generated. A standard HTK recogniser is used as the back-end in this paper. All the results shown in this paper, except the last table, uses DCT as the linear transform. The last table presents the results of the experiments that show the improvement in using a PCA transform instead of DCT. As noted earlier, in all the results, the average in the last row presents the average word level accuracy from 20 to 0 dB, as defined in the standardisation requirements [1].

### 3.2 Baseline Results

The baseline configuration consists of a standard MFCC front-end, with 13 static coefficients and their delta and delta-delta coefficients, making a total of 39 coefficients. The recognition results (word level accuracy) for the baseline are summarised in Table 1. The baseline average (from 20 to 0 dB) result over four different noise types is 85.6%.

<table>
<thead>
<tr>
<th>SNR (dB)</th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>N4</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>98.80</td>
<td>98.31</td>
<td>98.24</td>
<td>98.67</td>
<td>98.51</td>
</tr>
<tr>
<td>20</td>
<td>97.26</td>
<td>96.19</td>
<td>97.61</td>
<td>98.33</td>
<td>97.35</td>
</tr>
<tr>
<td>15</td>
<td>96.19</td>
<td>93.47</td>
<td>96.84</td>
<td>98.15</td>
<td>96.16</td>
</tr>
<tr>
<td>10</td>
<td>92.35</td>
<td>87.3</td>
<td>94.09</td>
<td>97.25</td>
<td>92.75</td>
</tr>
<tr>
<td>5</td>
<td>80.29</td>
<td>72.07</td>
<td>85.74</td>
<td>94.54</td>
<td>83.16</td>
</tr>
<tr>
<td>0</td>
<td>41.45</td>
<td>50.54</td>
<td>61.56</td>
<td>80.01</td>
<td>58.39</td>
</tr>
<tr>
<td>Average</td>
<td>81.51</td>
<td>79.91</td>
<td>87.17</td>
<td>93.66</td>
<td>85.56</td>
</tr>
</tbody>
</table>

**Table 1** Baseline word accuracies for four different noise types (N1, N2, N3, N4) at different SNRs.

### 3.3 Multi-Resolution Experiments

Experiments were then conducted to determine the improvement in performance when the full-band features were combined with sub-band features. The four sub-bands in this block were defined as follows:

- **Band 1**: 0–0.9 kHz
- **Band 2**: 0.9–1.8 kHz
- **Band 3**: 1.8–2.8 kHz
- **Band 4**: 2.8–4 kHz

Each sub-band produced 7 static features with their time derivatives, giving a feature vector dimension of 84. In the combined approach, the features from the sub-band and full-band were combined to produce a total feature vector size of 123 coefficients. The results with these three different experiments are summarised in Table 2.

It is seen from the table that the sub-band based speech recogniser produced worse word level accuracy than a full-band based recogniser. However, the use of full-band information together with the sub-band based features produced better performance than using either of them alone, with a word error rate reduction of approximately 5.3% compared to that of using full-band features alone. This clearly shows the importance of combining the sub-band based information with full-band features. We noticed that the sub-band features also improved the performance in clean conditions compared to the full-band features. It should be noted here that all the methods, whose results are given in Table 2, use cepstral domain feature vector normalisation [11].

We also conducted some experiments to find the optimum number of sub-bands and the number of sub-band blocks that could be used together with the full-band features. Table 3 gives the average results with different combinations of the sub-bands and blocks as explained below.

<table>
<thead>
<tr>
<th>Test Set</th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>N4</th>
<th>Average WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>88.56</td>
<td>83.20</td>
<td>91.91</td>
<td>96.36</td>
<td>90.01</td>
</tr>
<tr>
<td>Set 2</td>
<td>88.41</td>
<td>83.63</td>
<td>91.41</td>
<td>96.57</td>
<td>90.01</td>
</tr>
<tr>
<td>Set 3</td>
<td>88.63</td>
<td>84.29</td>
<td>91.09</td>
<td>96.70</td>
<td>90.18</td>
</tr>
<tr>
<td>Set 4</td>
<td>88.37</td>
<td>84.35</td>
<td>90.29</td>
<td>96.46</td>
<td>89.87</td>
</tr>
<tr>
<td>Set 5</td>
<td>88.57</td>
<td>83.65</td>
<td>91.10</td>
<td>96.43</td>
<td>89.89</td>
</tr>
<tr>
<td>Set 6</td>
<td>88.32</td>
<td>84.06</td>
<td>91.06</td>
<td>96.56</td>
<td>90.00</td>
</tr>
</tbody>
</table>

**Table 2** Word accuracies for full-band, sub-band based features and the combination of both. The last column shows WER decrease wrt. the full-band results in the first row.

It can be seen from Table 3, that there is no increased benefit achieved by continuously adding different kinds of sub-band combinations beyond a certain point. Since the amount of training data was always the same, it is possible that higher feature vector dimension caused this performance degradation. Obviously, one needs to tune the front-end algorithm so that feature vector dimension matches the amount of training data. With this database, the best results are obtained with a combination of full-band and an additional block containing 4
sub-bands.

3.4 PCA Experiments

Typically, the log-compressed mel-filtered coefficients are highly correlated. A DCT transform is commonly used in speech recognition for de-correlating the log-spectral features. This is very important as it allows us to use diagonal covariance matrices when modelling the state probability distributions in HMMs. Our belief is that the de-correlation process can be improved by using a transform which better takes into account the characteristics in the full-band as well as in different sub-bands.

A data-driven PCA transform was therefore computed for both full-band and various frequency sub-bands. Because PCA is a data-dependent transform, it is important that the transformation coefficients are estimated from a different speech database used for testing, in order to avoid database dependence.

Table 4 shows the results of the experiments, illustrating the advantage in using a PCA transform instead of DCT for computing the cepstral coefficients. Here, a multi-resolution front-end with one additional block containing 4 sub-bands, which provided the best results, has been used. A PCA transform, pre-computed during the training phase using a large number of utterances of the TIMIT database, is used to produce the cepstral coefficients. The transform basis functions are selected to be the eigen-vectors corresponding to the 13 greatest eigen-values of matrix C, which is computed as

\[ C = \sum_{k=1}^{K} f_o f_o^T, \]

where \( f_o \) stands for the vector of spectral values. It is seen that the use of PCA transform provided a 5% relative increase in word level accuracy compared to that of using DCT.

<table>
<thead>
<tr>
<th>Type</th>
<th>N1</th>
<th>N2</th>
<th>N3</th>
<th>N4</th>
<th>Aver.</th>
<th>WER decr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCT</td>
<td>88.63</td>
<td>84.29</td>
<td>91.09</td>
<td>96.70</td>
<td>90.18</td>
<td>31.99</td>
</tr>
<tr>
<td>PCA</td>
<td>90.07</td>
<td>84.78</td>
<td>91.51</td>
<td>96.80</td>
<td>90.79</td>
<td>36.22</td>
</tr>
</tbody>
</table>

Table 4 Comparison of word level accuracies for multi-resolution front end with DCT and PCA.

4. CONCLUSIONS

We have proposed a new approach to noise robust feature extraction in this paper. The combination of conventional full-band MFCC features with the sub-band based features is shown to improve the recognition performance compared to the use of either of the two feature sets individually. It is clear that the use of sub-band based features increases the noise robustness of the system as noise tends to corrupt only parts of the spectrum. Independent processing of the sub-band features does not spread the noise across the different cepstral coefficients. We have shown here that by neglecting the information from the full-band, we are losing the information related to the correlation between different sub-bands. It was also noticed in our experiments that the sub-band based features showed better performance even in clean environment, compared to the use of only full-band features.

The use of PCA instead of DCT also improved the performance of our noise robust front-end by around 5%. We combined the independent information from the sub-bands and the full-band at the feature level, rather than at the back-end, thereby avoiding computationally intensive back-end structures. Our experimental results with the new approach produced 36.22% overall decrease in word error rate compared to conventional MFCC front-end, with improvement occurring in all noise conditions.

5. REFERENCES

10. “ETSI ES 201 108 v1.1.2: Speech Processing, Transmission and Quality aspects (STQ); Distributed Speech Recognition; Front-end feature extraction algorithm; Compression algorithm”, ETSI standard, April-2000.